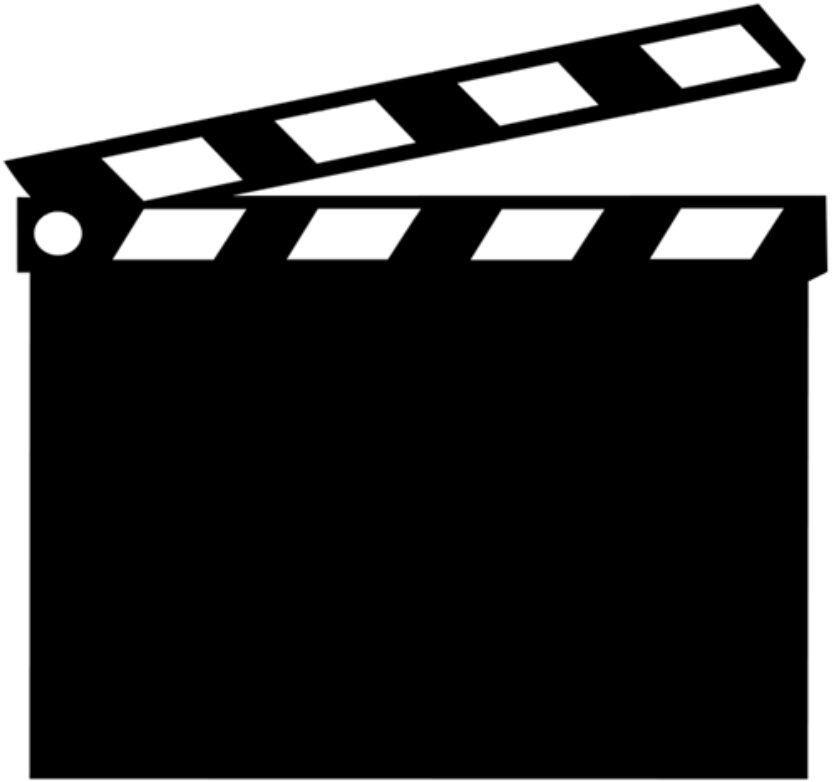
C018SA-W2-S7



SEMAINE 2 : Indexation

1. Introduction
2. Hiérarchie de mémoire
3. Fichiers indexés
4. Arbre-B
5. Hachage
6. Hachage dynamique
7. **Multi-hachage**

Filtre de Bloom

- 2 machines Paris et Lille
  - ✓ ensemble d'entiers A_p à Paris
  - ✓ ensemble d'entiers A_L à Lille
- On veut calculer $A_P \cap A_L$
- On envoie A_P (plus petit) à Lille
- On peut mieux faire ?

Filtre de Bloom – suite

- A Paris
  - on prend un tableau Bloom de N bits tous à zéro
  - On utilise k fonctions de hachage $h_1 \ldots h_k$
  - Pour chaque i dans [1,k] et chaque m dans A_p

    On met $Bloom[h_i(m)] = 1$
- On envoie le tableau *Bloom* à Lille

Remarque : tout objet m dans A_P satisfait

$$\forall i \in [1..k],\ Bloom[h_i(m)] = 1$$

Filtre de Bloom – suite

- À Lille
  - On calcule les *candidats*

$$S = \{\, m \in A_L \mid \forall\, i \in [1..k],\ Bloom[h_i(m)] = 1 \,\}$$

- On envoie S à Paris
  - ✓ $A_P \cap A_L \subseteq S$
- À Paris on calcule $A_P \cap S$
- Espoir : $|\, Bloom \,| + |\, S \,| < |\, A_P \,|$

Faux positifs

Chaque objet m dans A_P satisfait

$$\forall i \in [1..k],\ Bloom[h_i(m)] = 1$$

Chaque objet *de* $A_P \cap A_L$ satisfait donc cette propriété

- ✓ C'est ce qui fait que l'algorithme est correct

Certains objets de $A_L - A_P$ satisfont aussi la propriété

On les appelle les faux positifs

On les envoie pour rien

Filtre de Bloom – fin

- Choix à faire
  - La taille *N* de la table
  - Le nombre de fonctions de hachage

Pour minimiser les communications, il faut minimiser

$$| Bloom | + | S |$$

- N trop grand : $| Bloom |$ trop coûteux
- N trop petit : $| S |$ trop coûteux
- Pour un N donné, il faut trouver le meilleur k

Jointure par filtre de Bloom

- 2 machines Paris et Lille
  - ✓ Relation R_p[BA] à Paris et R_L[AC] à Lille
- On veut calculer $R_P \bowtie R_L$
  - ✓ En supposant R_P bien plus petit que R_L
- On envoie $Bloom(\Pi_A(R_P))$ à Lille
- A Lille, on calcule les candidats

$$S = \{ t \mid t \in R_L, \forall i \in [1..k], Bloom[h_i(\Pi_A(t))] = 1 \}$$

  - ✓ Peut-être des faux positifs
- A Paris $R_P \bowtie S$

MERCI